**Public health basis of screening**

**Framework:**

1. Definition
2. Difference between screening and diagnosis.
3. Aims of screening
4. Criteria for screening
5. Types of screening
6. Bias in screening
7. Designing and evaluating a screening programme.

**Definition:**

Screening is the presumptive identification of unrecognized disease or defect by application of tests, examinations, or other procedures that can be applied rapidly in apparently healthy population. (U.S. Commission on chronic illness, 1957)

* Considered as a preventive care function (secondary prevention), although some consider it as a logical extension of health care.
* It however differs from periodic health examination in the following respects-
  + Wider application,
  + Relatively inexpensive,
  + Time-effective for physicians.

**Difference between screening and diagnostic tests:**

|  |  |  |
| --- | --- | --- |
|  | **Screening tests** | **Diagnostic tests** |
| 1. | Done on apparently healthy population to detect potential cases or indicators. | Done on those with signs of a disease to establish presence or absence of disease. |
| 2. | Applied to a community or group of people | Applied to individuals |
| 3. | Based on one criterion or cut-off point | Based on evaluation of a number of evidences like symptoms, signs, and investigations. |
| 4. | Generally less accurate and relatively less expensive | More accurate but also more expensive |
| 5. | Not a basis of treatment | Forms a basis to initiate treatment |
| 6. | Initiative comes from the investigator | Initiative comes from a patient. |
| 7. | Simple, acceptable to patients and staff | May be invasive and cumbersome |
| 8. | Generally chosen towards high sensitivity not to miss potential disease | Chosen towards high specificity |

**Aim of screening:** To sort out those who have a disease or are at risk of having one from a group of apparently healthy persons and bring them under medical supervision and treatment.

Screening can not only be done for diseases but also for risk factors e.g. screening for high BP as a risk for stroke, screening for cholesterol levels as a risk factor for CVD. Also it is not always necessary to include some kind of technical procedure like X-rays or lab tests but simple procedures like BP assessment and certain criteria like MAST or CAGE can be also be successfully used for screening.

**Criteria for screening:**

1. **Wilson and Jungner classic screening criteria, WHO 1968**
2. The condition sought should be an important health problem
3. There should be an accepted treatment for patients with recognized disease.
4. Facilities for diagnosis and treatment should be available.
5. There should be a recognisable latent or early symptomatic stage.
6. There should be a suitable test or examination.
7. The test should be acceptable to the population.
8. The natural history of the condition, including development from latent to declared disease, should be adequately understood.
9. There should be an agreed policy on whom to treat as patients.
10. The cost of case-finding (including diagnosis and treatment of patients diagnosed) should be economically balanced in relation to possible expenditure on medical care as a whole.
11. Case-finding should be a continuing process and not a 'once and for all' project.
12. **Additional emerging screening criteria proposed over the past 40 years**
13. The screening programme should respond to a recognized need.
14. The objectives of screening should be defined at the outset.
15. There should be a defined target population.
16. There should be scientific evidence of screening programme effectiveness.
17. The programme should integrate education, testing, clinical services and programme management.
18. There should be quality assurance, with mechanisms to minimize potential risks of screening.
19. The programme should ensure informed choice, confidentiality and respect for autonomy.
20. The programme should promote equity and access to screening for the entire target population.
21. Programme evaluation should be planned from the outset.
22. The overall benefits of screening should outweigh the harm.

The criteria stated above can be transformed into some basic questions to be asked before starting a screening programme:

|  |  |
| --- | --- |
| The disease | Is it an important health problem? |
| Is the natural history well understood? |
| Is there a recognizable latent or early symptomatic stage? |
| Does early intervention improve the clinical/ public health outcome? |
| Screening test | Is the test valid? |
| Is the test simple, reliable and affordable? |
| Is the test acceptable to the patient and staff? |
| Diagnosis and treatment | Accessibility of diagnostic facilities |
| Is the treatment effective and accessible? |
| Is the overall programme cost-effective and sustainable |
| Does the benefit outweigh the harms from the programme? |

**What makes a disease appropriate for screening?**

1. **Magnitude and seriousness of disease:** The disease for which a screening is proposed should be an important public health problem in terms of its burden, morbidity or mortality. However, it is not necessary that an accurate magnitude of the burden will be required to initiate a programme. If a disease has very less burden or does not has a serious outcome it is probably better to focus the resources in other areas of greater need as the benefit from screening will only be little.
2. **Knowledge of natural history of disease and detectable preclinical phase:** Most diseases follow certain patterns, or natural histories, moving from the onset through a preclinical phase, and eventually on to the appearance of discernable signs and symptoms and advanced disease progression. There should be a known detectable preclinical phase during which the disease process is already present but the disease is not yet apparent clinically as it would be possible to decide precisely when a screening test should be applied in order to achieve maximum benefit and minimal overutilization of resources. In its absence, the screening programme would fail to find many undiagnosed cases. In the natural course of a disease there is a point beyond which cure may not be possible called ‘*critical point’*. If the detectable stage does not come early in the natural course of the disease before the critical point, or if the screening method is unable to identify cases before it is outside the range of current therapy, the utility of screening will be questionable as the treatment would not offer any considerable benefit over the usual time of diagnosis and treatment initiation. It is also necessary that, when untreated, the disease should progress to a clinical disease phase rather than remaining indefinitely without signs. Screening for disease that would never have progressed to the clinical phase is likely to increase the burden of the disease and treating those individuals would not decrease the actual burden but will subject them to any associated adverse effects. Designing a programme which maximizes the detection of cases with good prognosis, but which in the absence of screening may be unlikely to progress, will waste resources.
3. **Availability of treatment for the condition:** To be an appropriate target of a screening program, it is essential that the disease is treatable or modifiable by early intervention. Mere diagnosis is likely to just increase distress both mentally and financially among those diagnosed and their families in the absence of any intervention that may well alter the course of the disease. An early treatment should also offer an improved survival and fewer long term problems than late treatment. It should also be ensured that adequate facilities for treatment and access to the treatment are available.

**What makes a test suitable for screening?**

For a test to be suitable for screening, it is important to fulfil the criteria of *repeatability*, *validity*, *acceptability.* It should also be*, safe, simple*, applicable *rapidly* and with ease, *cost-effective*, and offer a good *yield* of cases.

1. **Acceptability:** How acceptable the test is to the population being screened and to those screening them. This takes into account a number of issues pertaining to the procedure itself in addition to ethics, safety of tests, and impact of the results. In general it implies a non-invasive test with high validity.

Tests that are painful or embarrassing like per-rectal examination for colon cancers, vaginal examinations may be less acceptable in a screening camp. Few patients may even not come for a test for the mere fear of the pain associated with getting pricked for blood samples. Tests that expose to potential hazards like radiations also may be less acceptable to the population. The results of a screening test may also have an impact on the acceptability of screening e.g. the stigma associated with HIV may make the people sceptical about a screening for it. Similarly screening for cancers may also be associated with considerable distress if the patient is tested positive. The people must also understand and accept the potential impact not only of getting a positive diagnosis from, but also the chances and the implications of getting incorrect screening results.

1. **Reliability:** It is also called as repeatability / precision/ reproducibility of a test. It is the ability of a test to produce consistent test results when repeated on the same individual in similar conditions. It depends on three major factors-
   1. **Observer variation:** it is again of 2 types
      1. **Intra-observer variation:** if the same observer takes 2 or more observations in the same individual on different occasions and gets a different reading everytime. May be minimized by taking average of reading.
      2. **Inter-observer variation:** This is the difference between different observers in the same individuals at the same time.

Observer errors are common in interpretation of X-rays, ECG, BP, histopathological specimens etc. They may be minimized by standardization of procedures, intensive training, taking repeat measurements etc. These errors can however, not be eliminated completely.

* 1. **Biological variation:** Due to various biological reasons like timing and circumstances of test, there are variations in the observation of physiological variables like BP, blood sugar, serum cholesterol etc. Unlike observer variation they will have to be measured repeatedly over time to minimize.
  2. **Technical errors:** defective instruments, faulty calibrations or reagents etc.

1. **Validity:** Also called Accuracy. It refers to what the test correctly measures what it purports to measure i.e. the closeness with which the measured value agrees with the true value. The true value is often that determined by a test considered as “gold standard”. It determines whether a test is good enough to be used in screening for a particular disease. Validity of a test is measured by its sensitivity and specificity.

|  |  |  |
| --- | --- | --- |
|  | **Disease present** | **Disease absent** |
| **Test positive** | True positive (a) | False positive (b) |
| **Test negative** | False negative (c) | True negative (d) |

**True positives-** Those individuals tested positive and have the disease in reality.

**False positives-** Those individuals tested positive but do not have the disease in reality.

**True negatives-** Those individuals tested negative and do not have the disease in reality.

**False negatives-** Those individuals tested negative but have the disease in reality. These are difficult to determine in a test immediately as the test negatives are not justified to be put for retesting.

**Accuracy:** The proportion of all the test results, both positive and negative, those are correct.

Accuracy= (a+d) / (a+b+c+d)

* 1. **Sensitivity:** Itis defined as the ability of a test to detect all those with the disease in the screened population. This is expressed in percentage as the proportion of those with the disease in whom a screening test gives a positive result.

Sensitivity= a / (a+c)

As false negatives are difficult to determine, thus sensitivity is also difficult to determine initially in screening. A sensitive test should be chosen when the disease is severe or has high mortality without treatment e.g. TB, MI. A sensitive test is most helpful when it is negative.

* 1. **Specificity:** It is defined as the ability of a test to identify correctly those free of the disease in the screened population. This is expressed in percentage as the proportion of people free of the disease in whom the screening test gives a negative result.

Specificity= d / (b+d)

A specific test is chosen when false positive results can adversely affect the patient physically, emotionally, and financially e.g. cancers, HIV. A specific test is most useful when it is positive.

An ideal screening test should have 100% sensitivity and specificity but that rarely happens. Both sensitivity and specificity are inherent properties of a screening test i.e. it does not change with the characteristics of the population being screened. The sensitivity and specificity for a quantitative continuously distributed can be changed by chosen cut-off values. Too high a cut-off value will tend to diagnose lesser cases and hence decrease the sensitivity but will increase the specificity and vice versa.

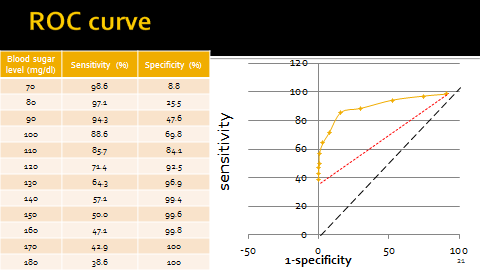
Many physiological variables are continuously distributed around the mean, conforming to a normal, or skewed normal, curve. Variables such as BP, blood cholesterol, blood sugar and intra-ocular tension-to give a few examples -all appear to favour a continuous distribution. The "diseased" part of the population occurs at the extreme end of the distribution curve and the "border-line" group in a population may be far greater than the "diseased".



If the distribution is bimodal, as in the case of some genetically transmitted characteristics, like phenylketonuria, the "border-line" group comprises a mixture of persons with and without the disease whose level of the variable falls within the same range (between A and B. On the other hand, if the distribution is unimodal the "border-line" group **will** comprise a homogeneous sample of persons, the question being whether the point between "disease" and "normality" should be set at C or D. If the cut-off point is set at B and D, it will indicate a highly specific test, whereas at point A and C it is highly sensitive. Generally in bimodal distribution the cut-off point is set at intersection of the diseased and normal curves ‘E’ as it will minimize both false negatives and false positives.

**Receptor operator characteristic (ROC) curve:** It is the graphical display of the how the proportions of true positives and false positives change for each of the possible pre-determined cut-off value in a test. We can plot a Receptor Operator Characteristics (ROC) curve by plotting true positivity rate (sensitivity) in Y-axis against false positivity rate (1-specificity) in X-axis. The better the test, the farther the curve is from the diagonal. The diagonal corresponds to the relationship between true-positive and false positive rates from a test yielding no information i.e. will not identify any undiagnosed case. ROC curves are particularly useful in comparing many tests for the same disease. The overall accuracy of a test can be estimated by the area under the curve (AUROC) - the larger the area, the better is the test. An AUROC of 1 corresponds to a perfect test and an AUROC of 0.5 corresponds to a test with no diagnostic value i.e. it lies on the diagonal. The shoulder of the curve is the usual point for setting cut-off. It has the following uses:

* Shows the trade-off between sensitivity and specificity i.e. with an increase in sensitivity the specificity decreases and vice-versa.
* To set a cut-off value for a test result.
* To compare the performance of different tests measuring the same outcome.



**Predictive value of a test:**

Although the two measures of sensitivity and specificity suffice in describing the validity of a test there are other measures which help in deciding whether a test should be used for screening or not. These are the predictive value of test. The predictive value of a test is not its inherent property as it depends on how good the test is (the sensitivity and specificity) and also the prevalence of disease. It is also sometimes called as the posterior or post-test probability. Predictive value can again be of two types-

1. **Predictive value of a positive test:** It is also called as Positive Predictive Value (PPV). It is the proportion of people actually having the disease among all those who have tested positive on the test. At individual level it answers the question that “If one tests positive on screening, what are the chances that he really has the disease?”

PPV = a / (a+b)

Usually the PPV of a test is quite low, thus it helps in determining which subset of population should be screened i.e. the high risk groups. A test with high PPV is more cost effective as it will record lesser proportion of individuals with false positive results and thus reduce the economic burden for confirmatory tests. PPV increases with prevalence of the disease being screened.

1. **Predictive value of a negative test:** It is also called as Negative Predictive Value (NPV). It is the proportion of people free of the disease among all those who have tested negative on the test. On the individual level “If one tests negative on screening, what are the chances that he is actually free of disease?” It helps to make the people understand that even though they test negative, there are still chances that the person may be having the disease, and thus explain the importance of follow-up testing at a later date. NPV has a limited role in determining the choice of test to be used. NPV increases with sensitivity.

NPV = d / (c+d)

N

**Likelihood ratios:**

Likelihood ratio is the probability of a particular test result in people with disease to probability of that result in people without disease.

It is an alternate way of describing the performance of a test. They can be used to calculate the probability of a disease after a positive or negative test result. It describes how many times more or less likely a test result is found in diseased than non-diseased. In a dichotomous test two types of likelihood ratio describe its ability to discriminate between diseased and non-diseased. Likelihood ratio can be used to calculate post-test odds-

Pre-test odds x likelihood ratio = post-test odds

Here, the pre-test odds is comparable to prevalence (pre-test probability), likelihood ratios to sensitivity / specificity, and post-test odds to PPV (post-test probability). It enables us to represent the results of a test as the degree of abnormality rather than mere presence or absence by enabling us to interpret the probability over an entire range of possible values. However, likelihood ratio is difficult to calculate as it requires conversion from probability to odds and back.

1. **Cost-effective:** The cost of screening is not only the cost of the procedure but of the entire follow-up process that is required after a positive result. The cost-effectiveness includes not only financial cost, but also non-financial costs both to the patients and the health personnel in terms of time, labour and inconvenience. To determine the cost-effectiveness of screening, issues like the false positive rate, predictive values of a test should be considered as they will lead to overtreatment. A high false negative rate also implies lesser cost-effectiveness as there is no point in engaging resources to a screening test which does not detect the diseased. Cost-effectiveness is helpful in guiding policy decisions.
2. **Yield:** It is the amount of previously unrecognized disease that is diagnosed as a result of the screening programme. It also depends on sensitivity, specificity and prevalence. A screening in high risk population usually leads to a higher yield of cases.

A screening test should also be *simple* enough to be administered by health care personnel other than physicians. Ideally, it should provide *rapid results* and be *inexpensive* enough to use on a broad scale. Another consideration is that there should be a mechanism for the follow-up of positive test results. Access to medical care affects the ability to obtain follow-up diagnostic tests if individuals screen positive, and to obtain effective medical treatment if the diagnostic tests are also positive. In situations in which follow-up testing and care are not available to members of a certain population group, screening tests are usually seen as not applicable for that group as these individuals would not benefit from merely being screened e.g. let us consider HIV testing. On one hand the screening test is simple enough these days by rapid diagnostic kits and also inexpensive but on the other hand it requires western blot to be confirmed. If individuals screening positive do not have easy access to a health facility offering the test, screening would only increase the stress and rather add to the patients suffering.

**Types of screening:**

Three types of screening have been described-

1. **Mass screening:** Mass screening is simply the screening of a whole population or community irrespective of the risk individuals have of contracting the disease. It has the advantage of covering the whole population but is less cost-effective and also more resource consuming.
2. **High risk or selective screening:** Screening applied selectively to high risk groups defined on the basis of epidemiological research e.g. HIV screening in all TB patients, VDRL testing in all pregnant women. For a screening program to be more effective it should be applied in a population which has a higher prevalence of the disease. This approach is much more cost effective, adds to the ‘program specificity’ and PPV of a test but the ‘program sensitivity’ decreases as cases from the unscreened population will be missed.
3. **Multiphasic screening:** It is the application of two or more tests in combination to the population being screened. The tests may be applied either in series or parallel.
   * When tests are applied in parallel/simultaneously it means that all individuals will go through the tests at the same time and if any of it comes positive the patient is considered positive. This approach increases the sensitivity and NPV of screening.
   * When tests are applied serially/sequentially it means that the individuals will undergo the first test (usually cheaper, less invasive) and if tested positive, will undergo the other screening test (usually costlier, invasive) and all tests must give positive result in order to label the patient as diseased. This approach tends to increase the specificity and PPV of a screening programme. Generally it is applied when one of the tests is expensive or hazardous. It is usually time consuming but cost-effective.

However, there is no evidence that multiphasic screening has any benefit to the population in terms of mortality and morbidity reduction, but it increases the cost of health services on the other hand.

**Bias in screening:**

Screening is subject to various kinds of biases:

1. **Referral / volunteer bias**: It is a kind of selection bias. In a screening it is likely that many people who volunteer to participate are healthier, more health conscious, and likely to comply with the advice. Such persons will have a better prognosis than others. However, it is also possible that a person volunteered because of positive family history or lifestyle characteristics which may or may not have a poor prognosis. It is often difficult to determine which way the bias affects the results of screening.
2. **Length bias**: This is also a type of selection bias which relates to the type of disease that is being screened. The natural history of disease varies in each individual. A disease may be rapidly progressing, with a shorter clinical and preclinical phase or slow progressing with longer preclinical and clinical phase. The latter form of disease usually has a better prognosis. Also persons with shorter preclinical form are likely to be missed by the screening and are more likely to be detected in between the screening intervals and are thus called as interval cases. Length bias usually relates with the concept that screening selectively identifies cases with a better prognosis and hence better survival. Such a bias would lead to a false perception of benefit in terms of improved survival even if in reality that may not be the truth.
3. **Lead-time bias**: It is an illusion of better survival only because of earlier detection. Lead time is the time difference between the actual detection of the disease by screening and its usual time of diagnosis in absence of screening. If in a disease early intervention does not benefit the patient in terms of survival, still it may be virtually visible that screening results in increased survival only because it lead to earlier diagnosis, while actually an early detection of the condition may have years of treatment and its associated side effects thus compromising the quality of life for that time.
4. **Overdiagnosis bias**: There may be two reasons for overdiagnosis bias-
   1. Often people initiating the screening programmes are over enthusiastic about it and tend to make incorrect or biased diagnosis (misclassification) to justify the initiative. Such cases will have a better prognosis and will thus show an incorrect but beneficial impact of screening in terms of improved survival due to the disease.
   2. In some cases it might be possible that a disease in preclinical phase would have never progressed to clinical stage or may also have regressed back to normal during the lifetime of an individual or that even though it would have progressed, it would have done so quite late in the lifetime to have a significant effect on the individual or the community. Screening will identify such cases and will only lead to falsely increase the burden of disease in the population.
   3. In a highly sensitive test with low specificity. For this kind of test a term ‘aggressive screening’ was coined by Moskowitz in 1976. However there is limited evidence of utility of such a test.

**Ethics in screening:**

1. Any screening program is initiated with a belief that it will on a whole improve the health status of the community and not necessarily every individual would be benefitted. Those planning to introduce a screening programme should be able to guarantee overall benefit to the community with a minimum risk that certain individuals may be disadvantaged by the programme.
2. Judicious use of limited resources especially in developing countries. in some cases screening could diminish the total level of health in a community by diverting the resources intended for routine care and other more pressing health problems. It should be ensured not to disturb the balance of health care system.
3. Informed consent – For screening, providing information about the test, its consequences, the diagnostic assessment, and treatments should also be presented, if a truly informed decision is to be made.
4. Psychological harm from false positives and unwarranted reassurance from false-negatives results.

**Evaluation of a screening programme:**

The evaluation of a screening programme helps us understand what benefit the people have gained through screening. There is very little justification for carrying or improving the screening process which offers no or minimal benefit to the population. However, some people consider screening as bound to be effective due to earlier intervention. The indicators which we look for in evaluation are:

1. **Operational measures**

* Number of people screened
* Proportion of target screened and number of times screened.
* Detected prevalence of the condition
* Total cost of programme
* Cost per case found
* Proportion of positive screenees brought to final diagnosis and treatment.
* Predictive value of the test

1. **Outcome measures**

* Mortality reduction
* Case-fatality rate reduction
* Percentage increase in cases detected early
* Complication reduction
* Prevention or reduction in recurrence
* Improvement in quality of life

**Summary: Designing a screening programme**

In designing a screening programme the following points have to be taken-

1. Disease to be screened- Burden, diagnostic workup, treatment facility.
2. Screening test- validity, reliability and applicability.
3. Determining the threshold/cut-off value –
   1. Higher cut-off gives specific but less sensitive test
   2. Lower cut-off gives sensitive but less specific test
4. Determining the population to be screened – Based on predictive values whether entire population (low risk) or selective (high risk) group is to be taken.
5. Counselling opportunities.
6. Arrangements for confirmatory test and treatment.
7. Evaluation of screening programme – Because of the deep-rooted belief among physicians that 'early diagnosis' of disease is beneficial, many regard screening as bound to be effective. However, for a number of reasons this is not necessarily the case thus necessitating evaluation of any screening intervention.
   1. The best study design to evaluate a screening programme is RCT. Although non-randomized trials, case control and cohort studies can also be used. For better evaluation mortality should be used as endpoint rather than survival.
   2. Referral bias, length bias, lead-time bias, and overdiagnosis bias.
8. Regular follow-up screening sessions.

References:

1. Mandelblatt JS, Cronin KA, Bailey S, Berry DA, de Koning HJ, Draisma G, et al. Effects of mammography screening under different screening schedules: model estimates of potential benefits and harms. Ann Intern Med. 2009 Nov 17;151(10):738–47.

2. Gulati R, Gore JL, Etzioni R. Comparative effectiveness of alternative prostate-specific antigen--based prostate cancer screening strategies: model estimates of potential benefits and harms. Ann Intern Med. 2013 Feb 5;158(3):145–53.

(1)(2)